Processing of multi-channel signals

The present invention relates to the processing of audio signals and, more particularly, the coding of multi-channel audio signals.

Parametric multi-channel audio coders generally transmit only one full-bandwidth audio channel combined with a set of parameters that describe the spatial properties of an input signal. For example, Fig. 1 shows the steps performed in an encoder 10 described in European Patent Application No. 02079817.9 filed November 20, 2002 (Attorney Docket No. PHNL021156).

In an initial step S1, input signals L and R are split into subbands 101, for example by time-windowing followed by a transform operation. Subsequently, in step S2, the level difference (ILD) of corresponding subband signals is determined; in step S3 the time difference (ITD or IPD) of corresponding subband signals is determined; and in step S4 the amount of similarity or dissimilarity of the waveforms which cannot be accounted for by ILDs or ITDs, is described. In the subsequent steps S5, S6, and S7, the determined parameters are quantized.

In step S8, a monaural signal S is generated from the incoming audio signals and finally, in step S9, a coded signal 102 is generated from the monaural signal and the determined spatial parameters.

Fig. 2 shows a schematic block diagram of a coding system comprising the encoder 10 and a corresponding decoder 202. The coded signal 102 comprising the sum signal S and spatial parameters P is communicated to a decoder 202. The signal 102 may be communicated via any suitable communications channel 204. Alternatively or additionally, the signal may be stored on a removable storage medium 214, which may be transferred from the encoder to the decoder.

Synthesis (in the decoder 202) is performed by applying the spatial parameters to the sum signal to generate left and right output signals. Hence, the decoder 202 comprises a decoding module 210 which performs the inverse operation of step S9 and extracts the sum signal S and the parameters P from the coded signal 102. The decoder further comprises a synthesis module 211 which recovers the stereo components L and R from the sum (or dominant) signal and the spatial parameters.

One of the challenges is to generate the monaural signal S, step S8, in such a way that, on decoding into the output channels, the perceived sound timbre is exactly the same as for the input channels.

Several methods of generating this sum signal have been suggested previously. In general these compose a mono signal as a linear combination of the input signals. Particular techniques include:

1. Simple summation of the input signals. See for example 'Efficient representation of spatial audio using perceptual parametrization', by C. Faller and F. Baumgarte, WASPAA'01, Workshop on applications of signal processing on audio and acoustics, New Paltz, New York, 2001.

2. Weighted summation of the input signals using principle component analysis (PCA). See for example European Patent Application No. 02076408.0 filed April 10, 2002 (Attorney Docket No. PHNL020284) and European Patent Application No. 02076410.6 filed April 10, 2002 (Attorney Docket No. PHNL020283). In this scheme, the squared weights of the summation sum up to one and the actual values depend on the relative energies in the input signals.

3. Weighted summation with weights depending on the time-domain correlation between the input signals. See for example 'Joint stereo coding of audio signals', by D. Sinha, European patent application EP 1 107 232 A2. In this method, the weights sum to +1, while the actual values depend on the cross-correlation of the input channels.

4. US 5,701,346, Herre et al discloses weighted summation with energy-preservation scaling for downmixing left, right, and center channels of wideband signals. However, this is not performed as a function of frequency.

These methods can be applied to the full-bandwidth signal or can be applied on band-filtered signals which all have their own weights for each frequency band. However, all methods described have one drawback. If the cross-correlation is frequency-dependent, which is very often the case for stereo recordings, coloration (i.e., a change of the perceived timbre) of the sound of the decoder occurs.

This can be explained as follows: For a frequency band that has a cross-correlation of +1, linear summation of two input signals results in a linear addition of the

signal amplitudes and squaring the additive signal to determine the resultant energy. (For two in-phase signals of equal amplitude, this results in a doubling of amplitude with a quadrupling of energy.) If the cross-correlation is 0, linear summation results in less than a doubling of the amplitude and a quadrupling of the energy. Furthermore, if the cross-correlation for a certain frequency band amounts –1, the signal components of that frequency band cancel out and no signal remains. Hence for simple summation, the frequency bands of the sum signal can have an energy (power) between 0 and four times the power of the two input signals, depending on the relative levels and the cross-correlation of the input signals.

The present invention attempts to mitigate this problem and provides a method according to claim 1.

If different frequency bands tended to on average have the same correlation, then one might expect that over time distortion caused by such summation would average out over the frequency spectrum. However, it has been recognised that, in multi-channel signals, low frequency components tend to be more correlated than high frequency components. Therefore, it will be seen that without the present invention, summation, which does not take into account frequency dependent correlation of channels, would tend to unduly boost the energy levels of more highly correlated and, in particular, psycho-acoustically sensitive low frequency bands.

The present invention provides a frequency-dependent correction of the mono signal where the correction factor depends on a frequency-dependent cross-correlation and relative levels of the input signals. This method reduces spectral coloration artefacts which are introduced by known summation methods and ensures energy preservation in each frequency band.

The frequency-dependent correction can be applied by first summing the input signals (either summed linear or weighted) followed by applying a correction filter, or by releasing the constraint that the weights for summation (or their squared values) necessarily sum up to +1 but sum to a value that depends on the cross-correlation.

It should be noted that although the invention can be applied to any system where two or more two input channels are combined.

Embodiments of the invention will now be described with reference to the accompanying drawings, in which:

Figure 1 shows a prior art encoder;

Figure 2 shows a block diagram of an audio system including the encoder of Figure 1;

Figure 3 shows the steps performed by a signal summation component of an audio coder according to a first embodiment of the invention; and

Figure 4 shows linear interpolation of the correction factors m(i) applied by the summation component of Figure 3.

According to the present invention, there is provided an improved signal summation component (S8') , in particular for performing the step corresponding to S8 of Figure 1. Nonetheless, it will be seen that the invention is applicable anywhere two or more signals need to be summed. In a first embodiment of the invention, the summation component adds left and right stereo channel signals prior to the summed signal S being encoded, step S9.

Referring now to Figure 3, in the first embodiment, the left (L) and right (R) channel signals provided to the summation component comprise multi-channel segments m1, m2... overlapping in successive time frames t(n-1), t(n), t (n+1). Typically sinusoids, are updated at a rate of 10ms and each segment m1, m2... is twice the length of the update rate, i.e. 20ms.

For each overlapping time window t(n-1),t(n),t(n+1) for which the L,R channel signals are to be summed, the summation component uses a (square-root) Hanning window function to combine each channel signal from overlapping segments m1,m2... into a respective time-domain signal representing each channel for a time window, step 42.

An FFT (Fast Fourier Transform) is applied on each time-domain windowed signal, resulting in a respective complex frequency spectrum representation of the windowed signal for each channel, step 44. For a sampling rate of 44.1kHz and a frame length of 20ms, the length of the FFT is typically 882. This process results in a set of K frequency components for both input channels (L(k), R(k)).

In the first embodiment, the two input channels representations L(k) and R(k) are first combined by a simple linear summation, step 46. It will be seen, however, that this could easily be extended to weighted summation. Thus, for the present embodiment, sum signal S(k) comprises:

$$S(k) = L(k) + R(k)$$

Separately, the frequency components of the input signals L(k) and R(k) are grouped into several frequency bands, preferably using perceptually-related bandwidths (ERB or BARK scale) and, for each subband *i*, an energy-preserving correction factor m(*i*) is computed, step 45:

$$m^2(i) = \frac{\sum_{k \in i} \left\{ |L(k)|^2 + |R(k)|^2 \right\}}{2 \sum_{k \in i} |S(k)|^2} = \frac{\sum_{k \in i} \left\{ |L(k)|^2 + |R(k)|^2 \right\}}{2 \sum_{k \in i} |L(k) + R(k)|^2} \qquad \text{Equation 1}$$

which can also be written as:

$$m^2(i) = \frac{1}{2} \frac{\sum_{k \in i} \left\{ |L(k)|^2 + |R(k)|^2 \right\}}{\sum_{k \in i} |L(k)|^2 + \sum_{k \in i} |R(k)|^2 + 2\rho_{LR}(i) \sqrt{\sum_{k \in i} |L(k)|^2 \sum_{k \in i} |R(k)|^2}} \qquad \text{Equation 2}$$

with $\rho_{LR}(i)$ being the (normalized) cross-correlation of the waveforms of subband *i*, a parameter used elsewhere in parametric multi-channel coders and so readily available for the calculations of Equation 2. In any case, step 45 provides a correction factor m(*i*) for each subband i.

The next step 47 then comprises multiplying the each frequency component S(k) of the sum signal with a correction filter C(k):

$$S'(k) = S(k)C(k) = C(k)L(k) + C(k)R(k) \qquad \text{Equation 3}$$

It will be seen from the last component of Equation 3 that the correction filter can be applied to either the summed signal (S(k) alone or each input channel (L(k),R(k)). As such, steps 46 and 47 can be combined when the correction factor m(*i*) is known or performed separately with the summed signal S(k) being used in the determination of m(*i*), as indicated by the hashed line in Figure 3.

In the preferred embodiments, the correction factors m(*i*) are used for the center frequencies of each subband, while for other frequencies, the correction factors m(*i*) are interpolated to provide the correction filter C(k) for each frequency component (k) of a subband *i*. In principle, any interpolation function can be used, however, empirical results have shown that a simple linear interpolation scheme suffices, Figure 4.

Alternatively, an individual correction factor could be derived for each FFT bin (i.e., subband i corresponds to frequency component k), in which case no interpolation is necessary. This method, however, may result in a jagged rather than a smooth frequency behaviour of the correction factors which is often undesired due to resulting time-domain distortions.

In the preferred embodiments, the summation component then takes an inverse FFT of the corrected summed signal S'(k) to obtain a time domain signal, step 48. By applying overlap-add for successive corrected summed time domain signals, step 50, the final summed signal s1,s2... is created and this is fed through to be encoded, step S9, Figure 1. It will be seen that the summed segments s1, s2... correspond to the segments m1, m2... in the time domain and as such no loss of synchronisation occurs as a result of the summation.

It will be seen that where the input channel signals are not overlapping signals but rather continuous time signals, then the windowing step 42 will not be required. Similarly, if the encoding step S9 expects a continuous time signal rather than an overlapping signal, the overlap-add step 50 will not be required. Furthermore, it will be seen that the described method of segmentation and frequency-domain transformation can also be replaced by other (possibly continuous-time) filterbank-like structures. Here, the input audio signals are fed to a respective set of filters, which collectively provide an instantaneous frequency spectrum representation for each input audio signal. This means that sequential segments can in fact correspond with single time samples rather than blocks of samples as in the described embodiments.

It will be seen from Equation 1 that there are circumstances where particular frequency components for the left and right channels may cancel out one another or, if they have a negative correlation, they may tend to produce very large correction factor values $m^2(i)$ for a particular band. In such cases, a sign bit could be transmitted to indicate that the sum signal for the component S(k) is:

$$S(k) = L(k) - R(k)$$

with a corresponding subtraction used in equations 1 or 2.

Alternatively, the components for a frequency band i might be rotated more into phase with one another by an angle $\alpha(i)$. The ITD analysis process S3 provides the (average) phase difference between (subbands of the) input signals L(k) and R(k). Assuming that for a certain frequency band i the phase difference between the input signals is given by $\alpha(i)$, the input signals L(k) and R(k) can be transformed to two new input signals L'(k) and R'(k) prior to summation according to the following:

$$L'(k) = e^{jc\alpha(i)} L(k)$$
$$R'(k) = e^{-j(1-c)\alpha(i)} R(k)$$

with c being a parameter which determines the distribution of phase alignment between the two input channels ($0 \leq c \leq 1$).

In any case, it will be seen that where for example two channels have a correlation of +1 for a sub-band i, then $m^2(i)$ will be ¼ and so $m(i)$ will be ½. Thus, the correction factor C(k) for any component in the band *i* will tend to preserve the original energy level by tending to take half of each original input signal for the summed signal. However, as can be seen from Equation 1, where a frequency band *i* of a stereo signal includes spatial properties, the energy of the signal S(k) will tend to get smaller than if they were in phase, while the sum of the energies of the L,R signals will tend to stay large and so the correction factor will tend to be larger for those signals. As such, overall energy levels in the sum signal will still be preserved across the spectrum, in spite of frequency-dependent correlation in the input signals.

In a second embodiment, the extension towards multiple (more than two) input channels is shown, combined with possible weighting of the input channels mentioned above. The frequency-domain input channels are denoted by $X_n(k)$, for the k-th frequency component of the n-th input channel. The frequency components k of these input channels are grouped in frequency bands *i*. Subsequently, a correction factor $m(i)$ is computed for subband *i* as follows:

$$m^2(i) = \frac{\sum_n \sum_{k \in i} |w_n(k) X_n(k)|^2}{n \sum_{k \in i} \left| \sum_n w_n(k) X_n(k) \right|^2}$$

In this equation, $w_n(k)$ denote frequency-dependent weighting factors of the input channels n (which can simply be set to +1 for linear summation). From these correction factors m(i), a correction filter C(k) is generated by interpolation of the correction factors m(i) as described in the first embodiment. Then the mono output channel S(k) is obtained according to:

$$S(k) = C(k) \sum_n w_n(k) X_n(k)$$

It will be seen that using the above equations, the weights of the different channels do not necessarily sum to +1, however, the correction filter automatically corrects

for weights that do not sum to +1 and ensures (interpolated) energy preservation in each frequency band.